Data Center Energy Trends

- Data center electricity usage
  - Increased by 56% from 2005 to 2010
  - 1.1% to 1.5% total world electricity usage
  - 1.7% to 2.2% total US electricity
    - (Note: Includes impact of 2008 recession.)
    - (Note: 2x increase 2000 to 2005, below prediction.)
  - Source: Koomey 2011

The Consequence

- At current growth rate (2000-2005) in energy usage for data centers, will need 30 new coal-fired or nuclear power plants by 2015

<table>
<thead>
<tr>
<th>% of World CO₂ Emissions</th>
<th>0.3</th>
<th>0.6</th>
<th>0.8</th>
<th>1</th>
<th>470</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Centers</td>
<td>146</td>
<td>146</td>
<td>146</td>
<td>146</td>
<td>146</td>
</tr>
<tr>
<td>Airlines</td>
<td>170</td>
<td>170</td>
<td>170</td>
<td>170</td>
<td>170</td>
</tr>
<tr>
<td>Shipyards</td>
<td>146</td>
<td>146</td>
<td>146</td>
<td>146</td>
<td>146</td>
</tr>
<tr>
<td>Steel plants</td>
<td>146</td>
<td>146</td>
<td>146</td>
<td>146</td>
<td>146</td>
</tr>
</tbody>
</table>

Four-fold increase surpasses airline industry!

Source: Koomey 2011
Increasing Memory Demand

- Parallelism (core count)
- Larger & complex data sets
- More sophisticated applications
- Virtualization & consolidation

Today: 10's (to 100's) GB
Tomorrow: Terabyte and beyond???

Figure 1: Projected annual growth in number of cores and memory capacity.
The expected number of cores per socket (blue line) is growing at a faster rate than the expected DRAM capacity (orange line). On average, memory capacity per processor core is extrapolated to decrease 30% every two years.

More Memory

- Energy/power consumption shift

Terabyte in Buffered DRAM or DDR3 SDRAM
- 8GB: 125 DIMMs, 400W@DDR3, 1.25KW@FBDRAM
- Up to 4-10x more than already power hungry machines!

DRAM

- A long-time winner: Decades old!
  - Cost, power, performance trade-offs have favored it
  - Massive future capacity leads to a different outcome!

- Limitations to DRAM
  - Destructive reads: Must replace data after a read
  - Limited data retention: Periodic refresh
  - Susceptibility to errors: Charge can be disturbed
  - Scalability: Projections (ITRS) question below 22nm
The Wave Rolling In

- DRAM has long been the best choice until now...
- DRAM does offer advantages
  - Effectively unlimited write endurance (doesn't wear out)
  - Fast read/write (symmetric) latency
  - (And, of course, it's a commodity, here today, etc.)
- Can we use it judiciously? Just a little bit, please?
  - Combine with alternative technology
  - Small DRAM has reasonable energy, capacity
  - We've seen this before... SRAM cache vs DRAM?

---

US Patents Granted

For an "old technology", a dramatic change of events with tremendous interest!

Source: Lam, VLSI-TSA 2008

---

Alternative Memory Technology

<table>
<thead>
<tr>
<th>Technology</th>
<th>Read Speed</th>
<th>Write Speed</th>
<th>Cell Area</th>
<th>Endurance</th>
<th>Addressability</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM</td>
<td>20~50ns</td>
<td>20~50ns</td>
<td>6F²</td>
<td>10¹⁵</td>
<td>Yes</td>
</tr>
<tr>
<td>SRAM</td>
<td>~2ns</td>
<td>~2ns</td>
<td>1.4kF²</td>
<td>10¹¹~10¹⁵</td>
<td>Yes</td>
</tr>
<tr>
<td>NAND Flash</td>
<td>25ns</td>
<td>500ns</td>
<td>5F²</td>
<td>10¹⁰~10¹¹</td>
<td>No</td>
</tr>
<tr>
<td>STT-RAM</td>
<td>2ns</td>
<td>10ns</td>
<td>37~40F²</td>
<td>10¹²</td>
<td>Yes</td>
</tr>
<tr>
<td>PCM</td>
<td>30~50ns</td>
<td>~1ns</td>
<td>8~8F³</td>
<td>10⁸~10⁹</td>
<td>Yes</td>
</tr>
</tbody>
</table>
### Alternative Memory Technology

<table>
<thead>
<tr>
<th>Technology</th>
<th>Read Speed</th>
<th>Write Speed</th>
<th>Cell Area</th>
<th>Endurance</th>
<th>Addressability</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM</td>
<td>20~50ns</td>
<td>20~50ns</td>
<td>6F</td>
<td>$10^{12}$</td>
<td>Yes</td>
</tr>
<tr>
<td>SRAM</td>
<td>~2ns</td>
<td>~2ns</td>
<td>146F</td>
<td>$10^2$</td>
<td>No</td>
</tr>
<tr>
<td>NAND Flash</td>
<td>25ns</td>
<td>500ns</td>
<td>5F</td>
<td>$10^{10}$</td>
<td>No</td>
</tr>
<tr>
<td>STT-RAM</td>
<td>2ns</td>
<td>10ns</td>
<td>2.5~5F</td>
<td>$10^{12}$</td>
<td>Yes</td>
</tr>
<tr>
<td>PCM</td>
<td>30~50ns</td>
<td>~1ns</td>
<td>2~8F</td>
<td>$10^{10}$</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Fast, non-destructive reads: Nearing parity w/ DRAM**
Non-volatile, non-destructive, no refresh, low energy

**Density on par with DRAM, 2.5nm prototype**

### Alternative Memory Technology

<table>
<thead>
<tr>
<th>Technology</th>
<th>Read Speed</th>
<th>Write Speed</th>
<th>Cell Area</th>
<th>Endurance</th>
<th>Addressability</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM</td>
<td>20~50ns</td>
<td>20~50ns</td>
<td>6F²</td>
<td>10¹⁵</td>
<td>Yes</td>
</tr>
<tr>
<td>DRAM</td>
<td>20~50ns</td>
<td>20~50ns</td>
<td>6F²</td>
<td>10¹⁵</td>
<td>Yes</td>
</tr>
<tr>
<td>NAND Flash</td>
<td>25ns</td>
<td>500ns</td>
<td>5F²</td>
<td>10¹⁵</td>
<td>No</td>
</tr>
<tr>
<td>STT-RAM</td>
<td>2ns</td>
<td>10ns</td>
<td>37~40F²</td>
<td>10¹²</td>
<td>Yes</td>
</tr>
<tr>
<td>PCM</td>
<td>30~50ns</td>
<td>~1us</td>
<td>2~8F²</td>
<td>10¹⁰</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- **Write performance limited**
- Relatively slow bit cell writes but no block wear required like Flash
- Multiple write rounds of bit groups leading to 1us (Numonyx prototype)

Density on par with DRAM, 2.5nm prototype

Fast, non-destructive reads: Nearing parity w/DRAM

---

### Alternative Memory Technology

<table>
<thead>
<tr>
<th>Technology</th>
<th>Read Speed</th>
<th>Write Speed</th>
<th>Cell Area</th>
<th>Endurance</th>
<th>Addressability</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM</td>
<td>20~50ns</td>
<td>20~50ns</td>
<td>6F²</td>
<td>10¹⁵</td>
<td>Yes</td>
</tr>
<tr>
<td>DRAM</td>
<td>20~50ns</td>
<td>20~50ns</td>
<td>6F²</td>
<td>10¹⁵</td>
<td>Yes</td>
</tr>
<tr>
<td>NAND Flash</td>
<td>25ns</td>
<td>500ns</td>
<td>5F²</td>
<td>10¹⁵</td>
<td>No</td>
</tr>
<tr>
<td>STT-RAM</td>
<td>2ns</td>
<td>10ns</td>
<td>37~40F²</td>
<td>10¹²</td>
<td>Yes</td>
</tr>
<tr>
<td>PCM</td>
<td>30~50ns</td>
<td>~1us</td>
<td>2~8F²</td>
<td>10¹⁰</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- **Write performance limited by individual bit and group of bits**
- Limited write cycles but better than Flash

Density on par with DRAM, 2.5nm prototype

Fast, non-destructive reads: Nearing parity w/DRAM

---

### Alternative Memory Technology

<table>
<thead>
<tr>
<th>Technology</th>
<th>Read Speed</th>
<th>Write Speed</th>
<th>Cell Area</th>
<th>Endurance</th>
<th>Addressability</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM</td>
<td>20~50ns</td>
<td>20~50ns</td>
<td>6F²</td>
<td>10¹⁵</td>
<td>Yes</td>
</tr>
<tr>
<td>DRAM</td>
<td>20~50ns</td>
<td>20~50ns</td>
<td>6F²</td>
<td>10¹⁵</td>
<td>Yes</td>
</tr>
<tr>
<td>NAND Flash</td>
<td>25ns</td>
<td>500ns</td>
<td>5F²</td>
<td>10¹⁵</td>
<td>No</td>
</tr>
<tr>
<td>STT-RAM</td>
<td>2ns</td>
<td>10ns</td>
<td>37~40F²</td>
<td>10¹²</td>
<td>Yes</td>
</tr>
<tr>
<td>PCM</td>
<td>30~50ns</td>
<td>~1us</td>
<td>2~8F²</td>
<td>10¹⁰</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- **Write performance limited by individual bit and group of bits**

Density on par with DRAM, 2.5nm prototype

Fast, non-destructive reads: Nearing parity w/DRAM

---

**Fast, non-destructive reads: Nearing parity w/DRAM**

**Write performance limited by individual bit and group of bits**

**Density on par with DRAM, 2.5nm prototype**

**Similar array structure/operation as DRAM: bit/byte addressability**

**Repeated writes lead to wear on bit cell**

**Write performance limited by individual bit and group of bits**

**Density on par with DRAM, 2.5nm prototype**

**Fast, non-destructive reads: Nearing parity w/DRAM**
**Alternative Memory Technology**

<table>
<thead>
<tr>
<th>Technology</th>
<th>Read Speed</th>
<th>Write Speed</th>
<th>Cell Area</th>
<th>Endurance</th>
<th>Addressability</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM</td>
<td>20~50ns</td>
<td>20~50ns</td>
<td>6F²</td>
<td>10¹⁰</td>
<td>Yes</td>
</tr>
<tr>
<td>DRAM</td>
<td>~2ns</td>
<td>~2ns</td>
<td>4F²</td>
<td>10¹³</td>
<td>Yes</td>
</tr>
<tr>
<td>NAND Flash</td>
<td>25ns</td>
<td>500ns</td>
<td>5F²</td>
<td>10¹⁰</td>
<td>No</td>
</tr>
<tr>
<td>STT-RAM</td>
<td>2ns</td>
<td>~10ns</td>
<td>37~40F²</td>
<td>10¹²</td>
<td>Yes</td>
</tr>
<tr>
<td>PCM</td>
<td>30~50ns</td>
<td>~1ns</td>
<td>2~8F²</td>
<td>10¹⁰</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Nearly ideal complement (maybe replacement?) for DRAM (scales, low standby power, bit addressable, fast reads)

**BUT... must find techniques to overcome limitations**

**PCM: The Fundamental Idea**

- Similar process as CD-R
- Chalcogenide (GST)
- Application of heat changes state of material
- Resistance associated with each state stores a bit
  - Crystalline (low, SET, 1)
  - Amorphous (high, RESET, 0)
- Operation
  - Write: Heat/cool
  - Read: Measure resistance

**PCM Read/Write Operations**

- **Read**
  - Measure resistance
  - Low: logic 1 (SET)
  - High: logic 0 (RESET)
  - Relatively fast
  - Power efficient
  - Non-destructive

- **Writes**
  - Slow bit writes: heating/cooling: 50ns – 150ns
  - Limited parallel bit writes: large programming current
  - Long latency: 1000ns
  - High write energy
  - Heat stress leads to failure, with limited endurance (10⁷)

Diagram/photo: Micron Technology
http://www.micron.com/innovations/pcm.html
Consequences of PCM

- **Asymmetric read/write latency and bandwidth**
  - Reads projected to reach parity with DRAM
  - Writes will remain slow due to heating/cooling

- **Wear-out and endurance management**
  - Integrated relatively near CPU leads to heavy usage
  - E.g., one write/second: PCM fails in 110 days
  - Memory will quickly fail without precautions

Nonvolatility

Reliability

Important, desirable properties. Most focus has been on making it work first, then find ways to exploit these properties.

---

Rethinking Main Memory for PCM

**Starting Point: DRAM Main Memory**

- Sandy Bridge

---

Hybrid Memory Archetype

**Conventional memory adapted to PCM**

- ** Essential idea**
  - Small DRAM combined with a large PCM
  - "Capacity, low standby power, write performance, endurance"

- **Capacity of change, high volume**
  - Accelerate DRAM PCM
  - DRAM write cache
  - DRAM write buffer

- **Small DRAM (single fast DIMM)**
  - Write performance
  - Write energy
  - Endurance
  - Capacity, standby power
DRAM read/write cache

- Phase-change Main Memory Architecture (PMMA)

- DRAM replacement
- Maintain same interfaces
- Connectivity components
- Isolate changes to mem ctrl

- PMMA acts as cache
- Accesses to main memory made through the cache
- Write performance
- Endurance management

PMMA

- System Agent
- Acts as controller to DRAM/PCM
- Hit: Check tags, access AEB
- Miss: Check tags, access PCM & AEB

PMMA
Request Controller

- Operates on pages (larger than cache block from CPU)
- Processes requests & allocates resources
  - Multiple outstanding requests
  - Page allocation & eviction (AEB)
- Map physical to device address
- Book keeping
  - Track resources used, including what is cached & where
  - Map physical address (PA) to PCM device address (DA)
- IFB: High speed memory buffers inflight pages (AEB/PCM)
**RC: Read Miss**

- Read B
  - AEB tag array
  - PCM pages B/C
- IFB Busy Bitmap
- PSM
- AEB cntrl
- FSM
- Read B: PCM to AEB
  - Make request, copy to IFB
  - Page B: PCM to AEB
  - Page C is clean

**RC: Read Miss w/o Writeback**

- Read B
  - AEB tag array
  - PCM pages B/C
- IFB Busy Bitmap
- PSM
- AEB cntrl
- FSM
- Read B: PCM to AEB
  - Make request, copy to IFB
  - Page B: PCM to AEB
  - Page C is clean

**RC: Read Miss w/o Writeback**

- Read B
  - AEB tag array
  - PCM pages B/C
- IFB Busy Bitmap
- PSM
- AEB cntrl
- FSM
- Read B: PCM to AEB
  - Make request, copy to IFB
  - Page B: PCM to AEB
  - Page C is clean
RC: Read Miss w/o Writeback

Read B

DRAM

IFB Busy Bitmap

Page B: PCM to AEB
Copy to AEB

Suppose evicted page, C, was dirty: Miss with eviction

Hand-off to DRAM to finish read

Active Request Queue

FSM

AEB tag array

Read B

PCM cntrl

Page B: C to PCM
Copy to AEB

IFB (Pages)

AEB cntrl

V

PAdr

n-way

Suppose evicted page, C, was dirty: Miss with eviction

Hand-off to DRAM to finish read

Active Request Queue

FSM

AEB tag array

Read B

PCM cntrl

V

PAdr

n-way
① Optimization: Page Partitioning

Sub-page is request unit
1x tag/map per page
Requested on demand
Presence/absence tracked

Asymmetric size

Requested on demand
Sub-page is request unit

Small dirty granularity
① Optimization: Page Partitioning

- Block transfer unit
- Smallest data transfer
- Fixed to PCM banks
- Higher priority requests pre-empt below blocks

② Optimization: CW + AEB bypass

- Critical block (word) first
  - Deliver block generating miss to CPU
  - Transfer remaining blocks on page
- AEB bypass
  - Inflight pages can service requests, if data available
  - Data delivered directly from AEB

③ Optimization: RWR

- PCM read-write-read (RWR)
  - RWR avoids writing unchanged blocks in sub-page
  - Read verify detects failed page
  - Failed write leads to spare allocation
3 Optimization: RWR

- PCM read-write-read (RWR)
  - RWR avoids writing unchanged blocks in sub-page
  - Read verify detects failed page
  - Failed write leads to spare allocation

<table>
<thead>
<tr>
<th>Read old block</th>
<th>Write block</th>
<th>Read new block</th>
</tr>
</thead>
<tbody>
<tr>
<td>blk'</td>
<td>blk'</td>
<td>blk</td>
</tr>
<tr>
<td>same</td>
<td>same</td>
<td>evicted dirty sub-page allocate spare</td>
</tr>
</tbody>
</table>

1. Read old block
2. Check for difference
3. If different, write block

4 Optimization: Endurance

- AEB eviction policy (N-chance) to minimize writes
- Non-uniform writes to memory
  - Uneven writes cause pages to fail before others
  - Failed page(s): memory is now broken
- Wear-leveling to uniformly distribute writes
  - Wear pages at same level
  - Pages will fail at approximately same time
- Spare capacity
  - Replace failed pages on-demand
PMMA Energy-Delay

• Compared to equivalent capacity in DRAM-only system (16GB, 4 core)
• PMMA small DRAM (speed optimized) with large PCM

E*D improved (small losses/gains are wins, e.g., bwaves)

256MB DRAM (224MB AEB+32MB meta) is good compromise

1024, 2048B page is good compromise tag area vs. locality

Small performance gain (~10%)
Inherently, not much better than DRAM
IFB + spatial locality + faster DRAM

E*D improved from PCM's low read power, smaller DRAM power, and filtering of writes at DRAM

E*D improved (small losses/gains are wins, e.g., bwaves)
- Poor spatial locality combined with large footprint.
- Brings in lots of pages, which are shortly evicted due to footprint.
- Lots of extra cost...

**Compromise:** Small $E \times D$ gain, with small pages and moderate sized AES (224 MB)

- $E \times D$ improved (small losses/gains are wins, e.g., bwaves)
- 256 MB DRAM (224 MB AES + 32 MB meta) is good compromise
- 1024B vs 2048B page trades tag/spare table vs. locality
Read-Write Page Partitioning

Results for AEB size 224 MB (+32MB meta data)
1024B best overall result but larger metadata storage
R/W page partitioning recoups losses from 2048B

Lifetime: Cumulative Impact

- Wear-leveling is essential to achieve 8 years
- 7-chance and RWR also have a large impact
Summary

- PCM architectures
  - DRAM complement for main memory?
  - Flash replacement
  - Memory + storage combination
- Current front-runners share essential idea
  - Small DRAM + Large PCM
- Endurance on the way to being solved?
- Write bandwidth and energy likely to persist